AMENDMENT OF SOLICITATION/MODIFICATION OF CONTRACT

OMB No. 0690-0008

1. CONTRACT ID CODE N/A

2. AMENDMENT/MODIFICATION NO. AMENDMENT 0001	3.EFFECTIVE DATE
4. REQUISITION/PURCHASE REQ. NO. NRMAE000000020	5.PROJECT NO. (If applicable)
6. ISSUED BY CODE: OFA611:WLV U.S. DEPARTMENT OF COMMERCE/NOAA ACQUISITION MANAGEMENT DIVISION ADP CONTRACTS BRANCH 1305 EAST WEST HIGHWAY, ROOM 7604 SILVER SPRING, MD 20910	7. ADMINISTERED BY CODE (If other than Item 6)
8. NAME AND ADDRESS OF CONTRACTOR (No. Street, County, State and ZIP Code) ALL OFFERORS Code: FACILITY CODE:	[X]9A. AMENDMENT OF SOLICITATION NO. 52-DDNR-0-90030 9B. DATED (See Item 11) MAR 28, 2000 [ ]10A. MOD. OF CONTRACT/ORDER NO. 10B. DATED (See Item 13)

11. THIS ITEM ONLY APPLIES TO AMENDMENTS OF SOLICITATIONS

[X] The above numbered solicitation is amended as set forth in Item 14. The hour and date specified for receipt of Offers [ ] is extended, [X] is not extended. Offerors must acknowledge receipt of this amendment prior to the hour and date specified in the solicitation or as amended by one of the following methods: (a) By completing Items 8 and 15, and returning 1 copy of the amendment; (b) By acknowledging receipt of this amendment on each copy of the offer submitted; or © By separate letter or telegram which includes a reference to the solicitation and amendment numbers. FAILURE OF YOUR ACKNOWLEDGMENT TO BE RECEIVED AT THE PLACE DESIGNATED FOR THE RECEIPT OF OFFERS PRIOR TO THE HOUR AND DATE SPECIFIED MAY RESULT IN REJECTION OF YOUR OFFER. If by virtue of this amendment you desire to change an offer already submitted, such change may be made by telegram or letter, provided each telegram or letter makes reference to the solicitation and this amendment, and is received prior to the opening hour and date specified.

12. ACCOUNTING AND APPROPRIATION DATA (If required)

NOT APPLICABLE

13. THIS ITEM APPLIES ONLY TO MODIFICATIONS OF CONTRACTS/ORDERS. IT MODIFIES THE CONTRACT/ORDER NO. AS DESCRIBED IN ITEM 14.

[ ] A.THIS CHANGE ORDER IS ISSUED PURSUANT TO: (Specify authority) _____ THE CHANGES SET FORTH IN ITEM 14 ARE MADE IN THE CONTRACT/ORDER NO. IN ITEM 10A.

[ ] B.THE ABOVE NUMBERED CONTRACT/ORDER IS MODIFIED TO REFLECT THE ADMINISTRATIVE CHANGES (such as changes in paying office, appropriation data,etc.) SET FORTH IN ITEM 14, PURSUANT TO THE AUTHORITY OF FAR 43.103(b).

[ ] C.THIS SUPPLEMENTAL AGREEMENT IS ENTERED INTO PURSUANT TO AUTHORITY OF:

[ ] D.OTHER (Specify type of modification and authority)

E. IMPORTANT: Contractor [ ] is not, [ ] is required to sign this document and return ____ copies to the issuing office.

14. DESCRIPTION OF AMENDMENT/MODIFICATION (Organized by UCF section headings, including solicitation/contract subject matter where feasible).

1. The purpose of this Amendment is to revise the Solicitation as follows:

SECTION B

NOTES TO OFFERORS, NOTE F - The first paragraph is changed to read as follows:

"(1) Evaluation of financing proposals will be based upon October 1, 2000, as the first invoice payment and will continue for thirty-six (36) months. If monthly prices are adjusted because of system upgrade(s), offerors need to specify the month the adjustment occurs and the amount."

[Continued on Page 3]

Except as provided herein, all terms and conditions of the document referenced in Item 9A or 10A, as heretofore changed, remain unchanged and in full force and effect.

15A. NAME AND TITLE OF SIGNER (Type or print)	16A. NAME AND TITLE OF CONTRACTING OFFICER (Type or print) WILLIAM L. VOITK Contracting Officer
15B. CONTRACTOR/OFFEROR _____ (Signature of person authorized to sign)	16B. UNITED STATES OF AMERICA _____ (Signature of Contracting Officer
15C. DATE SIGNED	16C. DATE SIGNED

SECTION C

o Section 4.2.1, Large Scale Cluster (LSC). Fourth to last bullet is changed to read as follows:

"The capability of running at least two copies of the Government's projected largest job when any single computer is unavailable for user jobs."
Page C-12 of the Solicitation should be removed and replaced with the enclosed Page C-12.

o Section C.4.2.8, Facilities, has been revised to include additional requirements. Page C-18 of the Solicitation should be removed and replaced with the enclosed Pages C-18 and C-18A.

o Section C.4.3.8, Facilities, has been added to the Desired Features. Page C-20 of the Solicitation should be removed and replaced with the enclosed Pages C-20 and C-20A.

SECTION H

o H.7, SITE PREPARATION, is changed to read as follows:

"H.7 SITE PREPARATION

The Contractor shall prepare the site at its own expense and in accordance with the equipment environmental specifications furnished in their proposal. Any alterations or modifications in site preparation which are directly attributed to incomplete or erroneous equipment environmental specifications provided by the Contractor which would involve additional expenses to the Government shall be made at the expense of the Contractor. Any delay in the installation date resulting from site alterations or modifications as specified in the paragraph above will not be charged to the Government. See Sections C.4.2.8, FACILITIES and C.4.10, FACILITIES DESCRIPTION AND REQUIREMENT, for detailed description and drawings of computer facility."
Page H-5 of the Solicitation should be removed and replaced with the enclosed Page H-5.

SECTION J

o Section J.3, Benchmark Instructions has been revised. Section J.3, Pages J-5 through J-26 should be removed and replaced with the revised Section J.3, Pages J-5 through J-28.

SECTION L

o L.6.1, Technical Proposals. Tab 6, LARGE SCALE CLUSTER (LSC), Item 7. Message Passing Within a Node, the formula in this paragraph is changed to read as follows:

"7. Message Passing Within a Node. For multiple-CPU nodes that run user applications, give the latency and effective bandwidth of point to point and non-blocked bi-directional message passing within a node. Give values for raw message sizes of 2 bytes, 20 bytes, 200 bytes, 2 KB, 20 KB, 200 KB, and 2000 KB, or the closest available data points. The effective bandwidth b_e is defined as

$$b_e = \frac{N}{T_l + \left(N + N_E\right)/b}$$

where N is the raw message size, N_E is the size of the header and trailer information added to the raw message for

purposes of transmission, T_l is the latency for sending the message, and b is the raw bandwidth. Describe the data paths taken and the sources of message-handling overhead. How do the available message passing interfaces differ in performance?"
Page L-12 of the Solicitation should be removed and replaced with the enclosed Page L-12.

o L.6.1, Technical Proposals. Tab 7, ANALYSIS CLUSTER (AC), Item 7. Message Passing Within a Node, the formula in this paragraph is changed to read as follows:

"7. Message Passing Within a Node. For multiple-CPU nodes that run user applications, give the latency and effective bandwidth of point to point and non-blocked bi-directional message passing within a node. Give values for raw message sizes of 2 bytes, 20 bytes, 200 bytes, 2 KB, 20 KB, 200 KB, and 2000 KB, or the closest available data points. The effective bandwidth b_e is defined as

$$b_e = \frac{N}{T_l + \left(N + N_E\right)/b}$$

where N is the raw message size, N_E is the size of the header and trailer information added to the raw message for purposes of transmission, T_l is the latency for sending the message, and b is the raw bandwidth. Describe the data paths taken and the sources of message-handling overhead. How do the available message passing interfaces differ in performance?"
Page L-17 of the Solicitation should be removed and replaced with the enclosed Page L-17.

NOTE F - Section L contains the clause INVITATION TO PROPOSE FINANCING TERMS (FAR 52.232-31)(OCT 1995). For evaluation purposes, assume the following:

(1) Evaluation of financing proposals will be based upon October 1, 2000, as the first invoice payment and will continue for thirt-six (36) months. If monthly prices are adjusted because of system upgrade(s), offerors need to specify the month the adjustment occurs and the amount.

(2) In accordance with Federal Acquisition Regulation 32.205(c)(4) the time value of proposal specified contract financing arrangements shall be calculated using an interest rate of 5.9% as specified in Appendix C to Office of Management and Budget (OMB) Circular No. A-94, “Benefit-Cost Analysis of Federal Programs; Guidelines and Discounts.”

NOTE G - In preparing the cost proposal, Offerors should keep in mind that only 94% of the funds available annually, beginning in FY 2001, may be dedicated to the HPCS components, as indicated in Section C.2.

- Completion of the LSC throughput benchmark in no more than 14400 seconds of wall clock time on the initial delivery of the LSC, configured as it will be run in production at GFDL
- At least one substantial upgrade to the sustained throughput of the LSC, as measured by an LSC throughput benchmark, during the base contract period
- Options to further enhance the LSC throughput after the base contract period
- A minimum of 144 GB of total memory on the initial delivery of the LSC
- A minimum of 256 MB/processor
- Options for at least 512 MB/processor and 1 GB/processor of memory
- A minimum of 3 dTB of formatted disk, exclusive of system , residing on a fault-tolerant disk subsystem
- A minimum sustained total I/O bandwidth to this disk of 4 GB/sec
- The ability to store files of up to 100 GB in size on this disk
- The ability to read file formats written by the AC without explicit library calls for data conversion from within the application.
- Failover capability for job queuing and scheduling

  When any set of resources in the LSC fails,
  - batch jobs using those resources are rerun without user intervention
  - only interactive sessions hosted on the failed resources are lost
  - users must continue to be able to login interactively
- The capability of running at least two copies of the Government's projected largest job when any single computer is unavailable for user jobs
- The capability of the LSC to operate and be repaired in degraded mode
- Full functionality when any part of the AC is halted or powered off
- An availability level of 96% on every computer in the LSC

C.4.2.2 Analysis Cluster (AC)

- An AC of two or more computers
- At least one substantial upgrade to the sustained throughput of the AC, as measured by an AC throughput benchmark, during the base contract period
- At the end of September 2001, assumption of the T94's maintenance payments or replacement of the T94 with an upgrade to the AC that provides computational performance at least equivalent to the T94, as measured by an AC throughput benchmark
- De-installation and disposal of the T94 if it is replaced

- All hardware and storage media used for backup to be provided by the Contractor
- Pre-delivery access within 30 days after award to systems similar to those proposed for the HPCS
- Technical support during this pre-delivery access period
- Acceptance testing that begins within 90 days of award

C.4.2.8 Facilities

- An initial HPCS with a power demand that does not exceed 675 kVA, the amount that is available during the parallel operation with all of the equipment under the current SGI/Cray contract. The available power will increase as SGI/Cray equipment is removed, subject to power information provided in C.10.2.
- The Contractor must assure that sufficient power capacity is available for the HPCS, the remaining legacy equipment from the SGI/Cray contract, and other Government equipment in the Computer Building over the life of the contract from the Computer Building substation and the PSE&G substation, as prescribed in C.4.10.2. The Contractor is responsible for any necessary upgrade of the Computer Building substation at no additional cost to the Government. Also, the Contractor will have the responsibility to assure that adequate power utility service is provided to the building.
- An initial HPCS with cooling requirements that do not exceed 3100 KBTU/hr, which is the total available cooling capacity (5000 KBTU/hr) minus the estimated cooling load indicated in Table 5. The cooling available to the HPCS will increase as SGI/Cray equipment is removed, subject to the cooling information provided in C.10.3.
- The Contractor must assure that sufficient cooling capacity is available to the Computer Building for the HPCS, the remaining legacy equipment from the SGI/Cray contract, and other Government equipment in the Computer Building over the life of the contract, as prescribed in C.4.10.3. The Contractor is responsible for any necessary upgrade of the chilled water plant at no additional cost to the Government, including any additional power upgrades that are needed.
- Replacement of the air-conditioning (blazer) units and associated equipment in the Computer Room, as described in C.4.10.3.
- An HPCS with floor space requirements that do not exceed the available space in the Computer Room, as described in C.4.10.5
- Replacement of raised flooring as described in C.4.10.5

- Recommendations for how to execute a transition strategy to the follow-on system if the Contractor needs to use more than half of the available floor space for the

HPCS |

- Design and construction of Operations and User Support Rooms adjacent to the Computer Room and associated additional Computer Room changes (including monitoring cameras), according to the specification outlined in C.4.10.6. | | |

C.4.3 Desired features

The following features are considered desirable on the proposed HPCS:

C.4.3.1 Large Scale Cluster (LSC)

- Binary compatibility of all processors
- The identical configuration of all computational nodes
- The ability to run the same OS level on all computational nodes and processors
- The ability to run two or more different OS levels simultaneously on the same computer on the LSC
- The ability for a single message-passing application to access all of the computational nodes on the LSC
- Total memory that scales linearly with throughput on systems that exceed the minimum throughput requirements
- Resources for interactive work that can be isolated from the batch production resources
- The ability to reassign interactive resources to the batch production jobs during non-primetime hours without a reboot of the entire LSC
- The ability to test OS and application software upgrades in isolation from the interactive and batch production resources on the LSC
- User login to a single hostname
- Failover to processors that are binary-compatible with and running the same OS level as the failed processors

C.4.3.6 Software

- System software on the LSC and AC listed as desired in C.4.8.1
- Resource management software on the LSC, AC, HSMS, and HFS listed as desired in C.4.8.2
- Programming environment software on the LSC and AC listed as desired in C.4.8.3
- X-windows applications software for the AC listed as desired in C.4.8.4

C.4.3.7 Reliability, Availability, and Support

- Software that allows users to restore /home files from the automated backup via a graphical interface
- Minimal impact of the automated backups on the network load

C.4.3.8 Facilities

- Sufficiently low HPCS cooling requirement that only one chiller is required to operate on most days
- Sufficiently compact total HPCS foot print so that no more than half of the available raised floor space is neededC.4.2.8 Facilities
- An HPCS that meets the requirements for power, cooling, and floor space discussed in C.4.10.2 through C.4.10.5
- Replacement of the air-conditioning units and associated equipment as described in section C.4.10.3
- Replacement of flooring as described in section C.4.10.5
- Design and construction of rooms adjacent to the Computer Room as outlined in C.4.10.6

The requirements and desirable features on the HPCS are described in more detail below.

C.4.4 High-Performance Computing

The HPCS shall provide high-performance computing resources for large-scale computing and analysis capabilities.

C.4.4.1 Large Scale Cluster (LSC)

Scalable supercomputing capabilities shall be provided by a Large Scale Cluster (LSC) of two or more computers (defined in C.6 as the maximum set of nodes that may be unavailable during the repair of any subset of those nodes). The Government desires binary compatibility of all processors and desires the identical configuration of all computational nodes within the LSC (computational nodes are used for numerical computation in the production workload, rather than, e.g., nodes dedicated primarily to I/O). The homogeneity of computational nodes will be evaluated. The ability to run the same OS level on all computational nodes and processors and the ability to run two or more OS levels simultaneously on the same computer in the LSC is desired. It is desirable that a single message-passing application be able to access all of the computational nodes on the LSC.

C.4.4.1.1 LSC performance

The LSC must provide a substantial increase in sustained throughput over that provided by GFDL's current Cray supercomputers described in C.3.1. Sustained throughput shall be measured by an LSC throughput benchmark (section J.3 of this RFP) comprised of concurrently-run parallelized codes sampled from GFDL's expected future workload. This throughput benchmark shall run in no more than 14400 seconds of wallclock time on the initial delivery of the LSC, configured as it will be run in production at GFDL. In addition, the scalability of the LSC shall be measured by a benchmark designed to reveal the performance and scaling characteristics of individual codes as they are executed on

(c) Automobile Liability. The Contractor shall have automobile liability insurance written on the comprehensive form of policy. The policy shall provide for bodily injury and property damage liability covering the operation of all automobiles used in connection with performing the contract. Policies covering automobiles operated in the United States shall provide coverage of at least $200,000 per person and $500,000 per occurrence for bodily injury and $20,000 per occurrence for property damage.

(d) Aircraft Public and Passenger Liability. When aircraft are used in connection with performing the contract, the Contractor shall have aircraft public and passenger liability insurance. Coverage shall be at least $200,000 per person and $500,000 per occurrence for bodily injury, other than passenger liability, and $200,000 per occurrence for property damage. Coverage for passenger liability bodily injury shall be at least $200,000 multiplied by the number of seats or passengers, whichever is greater.

H.7 SITE PREPARATION

The Contractor shall prepare the site at its own expense and in accordance with the equipment environmental specifications furnished in their proposal. Any alterations or modifications in site preparation which are directly attributed to incomplete or erroneous equipment environmental specifications provided by the Contractor which would involve additional expenses to the Government shall be made at the expense of the Contractor. Any delay in the installation date resulting from site alterations or modifications as specified in the paragraph above will not be charged to the Government. See Sections C.4.2.8, FACILITIES and C.4.10, FACILITIES DESCRIPTION AND REQUIREMENT, for detailed description and drawings of computer facility.

H.8 TECHNOLOGY SUBSTITUTION

H.8.1 OVERVIEW

All items (e.g., hardware, system applications software) and support services (maintenance, training, documentation, installation, and technical support services) shall be the most modern and cost-effective available at the time of delivery and installation. The contractor shall propose substitute items whenever the contractor or its subcontractor is offering replacement or substitutes for the components in question and the contractor offers the particular product to any of its commercial or Government customers. The Government may request that those items be substituted for comparable items originally offered. The Government reserves the right to accept or reject proposed substitutions.

J.3 Benchmark Instructions

J.3.1 Overview

In order to be considered for award, Offerors must successfully complete the benchmarks described below. The benchmarks may be obtained by following the instructions at http://www.gfdl.gov/hpcs/RFI/gfdl_bench.html. Vendors that have already completed and submitted a Benchmark Software Agreement need not do so again.

The Offeror must provide in tar/gzip format the source code used and the requested verification output for all aspects of the benchmark, as described in Sections J.3.2.2.3, J.3.2.3.3, J.3.3.2 and J.3.3.3, on 100MB Zip disk or ISO-9660 CDROM. All written responses and spreadsheets called for in these sections must be returned with the RFP response in printed form and digitally on 100MB Zip disk or ISO-9660 CDROM.

J.3.1.1 Source Code Changes

The Offeror may make changes to the compilation process and run script as necessary to accommodate their particular compilation and runtime environment(s).

Additionally, the Offeror may make changes to source code. However the Government requires that its applications be able to run on many different types of machines. Source code changes that reduce portability increase the costs of software maintenance and upgrades across multiple architectures. Therefore, certain types of code changes are preferred while others are discouraged. For the purpose of evaluating offerings, source code changes are divided into 4 Classes:

A. Modifications required to make a model run correctly, consistent with ANSI standard FORTRAN90 and C

B. Modifications to the program parallel communication

C. Modifications consistent with ANSI standard FORTRAN90 and C

D. All other modifications

Class A modifications are those required to allow a benchmark to run to completion correctly if, without such changes to source code, the benchmark will "fail" either by exiting prior to completion or producing incorrect answers. Class A modifications do not include any changes to source solely for performance.

Since there may be many causes for such changes (e.g. GFDL non-standard language usage within the application, work-arounds required for compiler bugs, etc), the Government

cannot state categorically that such modifications will not be evaluated without some sort of risk factor assigned. Still, it is the Government's desire to consider such changes as "essentially unmodified" code with no negative impact on evaluation.

Among the types of "changes" which will be taken as Class A are:

- Use of commercially supported libraries which are bid as part of the offering that requires no changes to benchmark source code or introduction of wrapper subroutines

- Compiler command lines with performance-specific options including, but not limited to, automatic parallelization

- Automatic parallelization and multitasking mediated through the operating system

- Use of commercially available and supported source pre-processors which are bid as part of the offering.

Class B modifications are source code changes either to the MPP library (mpp.F90, mpp_io.F90 and mpp_domains.F90) or to the direct use of MPI within an application (as within the HIM application). This includes use of communication libraries other than MPI. Such changes are encouraged though maintenance of MPI will still be required.

Class C modifications are limited to those which do not reduce code portability and which remain consistent with ANSI standard FORTRAN90 and C (it is acknowledged that the codes as they exist may already contain some ANSI non-compliant features). Performance is important and the Government is interested in performance-enhancing code modifications. However, resources to implement and maintain such changes are limited. Thus while a risk assessment will be made of any such changes, they are encouraged.

Among the types of changes taken to be Class C are:

- Use of commercially supported libraries which are bid as part of the offering

- Use of compiler "directives" within the source

Class D modifications are all those changes to application source not included in Classes A, B, or C. Such modifications reduce code portability and tend to make development and maintenance more difficult and costly. **Class D modifications are very strongly discouraged**.

All acceptable changes must produce output consistent with the verification provided as described with each benchmark.

As described in the instructions below, baseline performance numbers comprised of only Class A modifications will be required. MPI will be the required communication library for this baseline where a communication library is employed. MPI (or any other communication library) is clearly not applicable for systems which use compiler or operating system mediated AUTOMATIC parallelization for the baseline benchmark.

Offerors wishing to make code changes for evaluation must submit complete performance numbers for the entire test suite containing the code changes IN ADDITION to the baseline numbers. Having satisfied the baseline requirement, the Offeror is free to mix classes of changes. Offerors are cautioned, however, that a set of performance numbers and the associated changes will be evaluated as a single entity and accepted or rejected as such.

While it is highly desirable, it is not required that the Offeror reach minimum performance requirements based on Class A changes alone. However, Offerors are again cautioned that source changes associated with a set of performance numbers are assessed risk as a single entity.

J.3.1.2 Performance Data

Gathering of performance data is targeted to a system equivalent to that offered for the initial delivery. In this vein, the Test Systems on which the benchmarks are run and for which performance data is reported should be as close possible to the initial offered system. In general, any component of a Test System which is not the component proposed for the initial offered system will require the Government to make a risk assessment. The reasons for assigning risk will be clearly stated to each Offeror in their evaluation.

Still, the Government acknowledges that it may not be possible to use the offered system for either the RFP response or the LTD. Therefore two scenarios are provided to allow Offerors to respond to the RFP. Note that each scenario carries its own level of risk assessment.

Scenario A

Risk Assessment: *low to medium risk*

The Offeror shall develop performance data for the RFP response on a Test System with not less than 25% of the proposed number of processing elements, 25% of the computational performance, and 25% of the application memory. Further, a system at least as capable shall be used in the LTD.

A variation of this scenario is one where the Offeror has access to a Test System fulfilling at least the "25% criteria" for the purposes of providing the RFP response data, but may not have access to such a system for the LTD. In this event, the Offeror will provide output from Test Systems with no less than 25% the number of processing elements, 25% the

computational performance, and 25% application memory of the offered system. The Offeror will then run components of the benchmark on the systems which are available at LTD to verify aspects of the performance numbers provided.

In either case, for Test Systems less than 100% of the offered system, the Offeror shall define, document and demonstrate the scalability features which will allow the delivered system to meet the offered performance values of the full system at installation. Clearly the smaller the Test System and LTD systems the greater the need for extrapolation and the higher the associated risk. Conversely the level of risk assessed for this scenario declines as the Test System for which data is collected for RFP response AND demonstrated at LTD reaches 100% of the offered system. The burden of proof and associated risk is increased when the LTD system is less than the Test System used for the RFP response.

It is the Offeror's responsibility to develop, document and explain the extrapolation methodology. This may require data not called out in this RFP. It is the Offeror's responsibility to define and provide this data. The Offeror must detail all aspects of the extrapolation methodology, the supporting data and the demonstration methodology in the RFP response. The offer may be judged non-compliant if the extrapolation or demonstration methodology or the supporting data is determined to be unsuitable.

Scenario B

Risk Assessment: *highest risk*

The Offeror may extrapolate performance of the delivered system entirely from Test Systems meeting less than the 25% criteria described in Scenario A. Similarly, the LTD would take place on systems meeting less than the 25% criteria.

As with Scenario A, it is the Offeror's responsibility to develop, document and explain the extrapolation methodology. This may require data not called out in this RFP. It is the Offeror's responsibility to define and provide this data. The Offeror must detail all aspects of the extrapolation methodology, the supporting data and the demonstration methodology in the RFP response.

While not wishing to exclude a priori Offerors in this situation, the Government is highly skeptical that responses developed under Scenario B can be found compliant. Offerors following this approach accept an exceptional burden of proof.

These two scenarios cover the RFP response and pre-award Live Test Demonstration (LTD) phases. The only acceptable post-award LTD will be to successfully run the entire throughput suite at the performance level proposed by the successful Offeror and as described by the Acceptance Criteria section of this document. **Note that the software and hardware system configuration for this post-award LTD is required to be the same as that proposed for**

the initial production configuration at GFDL.

J.3.2 Large Scale Cluster (LSC) Benchmark

J.3.2.1 Overview

The LSC benchmark is comprised of 2 parts with the following goals:

i) Throughput Benchmark: A measurement of system performance under quasi-realistic GFDL workload and Offeror proposed runtime environment to be completed in a maximum wall clock time of 14400 seconds.

ii) Scaling Study: A measurement of application performance, scaling and resource requirements with respect to a given GFDL "experiment".

There are 3 applications (or 4 depending on how one counts; the FMS "bgrid" and FMS "spectral" may be viewed as separate applications utilizing the same infrastructure) and 14 experiments derived from these. The experiments will be tested as a throughput suite in i) and individually tested in ii). They are:

1. FMS Spectral Atmosphere T42L20 coupled to 2deg MOM3 ocean
2. FMS Spectral Atmosphere T42L20 coupled to 1deg MOM3 ocean
3. FMS Spectral Atmosphere T106L30
4. FMS Lo-resolution N45L20 coupled to 1deg MOM3 ocean
5. FMS Hi-resolution N90L30 coupled to 1deg MOM3 ocean
6. FMS Standard Atmosphere N45L160
7. FMS Development N30L40
8. FMS Atmosphere N30L40 with Tracers
9. FMS Hi-resolution N270L40 Atmosphere
10. MOM3 2deg L36 ocean
11. MOM3 1deg L50 ocean
12. MOM3 3deg L25 + tracers
13. MOM3 p5deg MESO
14. HIM p25deg MESO

All experiments are to be run in 64-bit, IEEE floating point precision.

The throughput suite has been constructed from a set of "job streams" which can be completed in the 372 - T90 processor equivalent hours available to the lab in a 12 hour period. The phrase “T90 processor equivalent” means the computational capability of a single processor on the GFDL T932 with respect to a given experiment segment, or the number of GFDL T3E processing elements (PEs) required to produce the equivalent performance. A "job stream" is the set of sequentially processed segments of a given experiment which

completes in the 12 hours.

It is important to note that throughout the benchmark instructions, a one to one, though not necessarily static, mapping of application processes to physical, application processors is assumed. For architectures where this is not the case, it is incumbent upon the vendor to document the distinction between the number of application processes and application processors. In this context, "application processors" means those processors with some part of the GFDL application running on them. This does not include auxiliary processors who's role is to provide specific support functions (such as communications assists). Auxiliary processors do need to be documented as part of the system configuration.

Many of the jobs utilize input files, some of which are rather large restart files. Further, restart files as well as other output files are written by the jobs. It is highly desirable that the vendor test and demonstrate movement of these files from and to the data archive and the file system in which the experiment segment will be run as is performed by the scripts which accompany the individual experiments. It is assumed that the data would be on the "spinning disk" portion of the archive (i.e. there is no intent that retrieval from tape storage be part of this benchmark). The Government is specifically interested in tests which move data through the same software and hardware interconnect between "archive" and "LSC" as will be proposed by the Offeror.

It is understood that an offered HSMS may transfer data directly between the HSMS nearline tier and LSC disk (i.e. does not require transferring data from archive staging disk to a runtime directory on the LSC). In this case, the Offeror clearly should not introduce artificial file transfers into the benchmark.

J.3.2.2 LSC Throughput Benchmark

J.3.2.2.1 General Comments

The Throughput Benchmark should be performed within the following framework.

The queuing and scheduling software being proposed for the installed system should be used for the Throughput Benchmark. The Government acknowledges that the details of queues and scheduling used at install and after will likely be an evolving process. Still, based on the description of lab processing activities as well as the job stream for the Throughput Benchmark, the Offeror should construct queues and scheduling which may be generalized to be used by GFDL at install. Queue and scheduling structures which appear to be specialized merely to optimize throughput of the particular job stream of the benchmark fail generality and will be penalized. It is incumbent upon the Offeror to convince the Government that the queue structures and scheduling used during this throughput test meet the requirements of generality and extensibility.

Operators may not intervene to specify or alter the number of processing elements on which a job is running. If a job is paused or migrated in any way by the (human) operator, a description of the reason why and what was done must be provided.

The Offeror must not forget to include the time required for the file staging and storage to archive disk in extrapolations to the offered system. The Offeror is reminded that the 14400 second maximum throughput time for this benchmark on the installed system includes all file transfers necessary to and from "archive disk" and "application runtime" file systems, but does not include retrieval of files from tape storage to "archive disk".

By way of definition, a "job stream" is defined as the set of job segments totaling 12 hours of T90 processor equivalent time which completes an experiment. For the Throughput Benchmark, there are a total of 31 "job streams" comprised of the 14 different experiments.

It should be understood that three job streams have been combined into a single experiment for HIM (that is, the total run length of HIM has been specified such that it would take 36-T90 equivalent processor hours to complete). This will require that HIM be run on a proportionately larger number of processors than other experiments. As each experiment has 2 segments, there are a total of 29x2=58 experiment segments to be run for the Throughput Benchmark. Each of the 2 experiment segments must be completed sequentially.

Segments within a job stream are constrained as follows. The only changes to the job script allowed between segment submission within a given job stream (*e.g*. a single MOM3 p5deg stream) are changes required for starting from a restart file after the run from the 0 time step has been completed (where applicable).

PE specification may be different BETWEEN streams of the same application. It is the PE specification WITHIN a stream which must remain invariant. In particular, the specification for the number of processing elements employed by the job may NOT be changed in either the script or on the submission command line between submissions within a stream. Also as mentioned above, the number of PEs may not be specified by operator intervention. This does not mean that a stream segment must run on the same number of processors for each submission. But it does imply that selection from a range or set of possible processor configurations for a given run must be specified within the job script and thereafter handled automatically by software.

The reasons for the constraints on specifying PEs within a job stream are as follows. Assuming that a given application will run on a range of PE configurations, scientists at GFDL will not know a priori the processor configuration on which to run a given job that will optimize turn-around time and resource utilization efficiency at the time the job finally starts. Therefore, the scientists will choose a PE configuration (or a range or set of PE configurations if the queuing software allows) at the time of submission which they feel meets their requirements for turn around time and resource utilization efficiency. This choice of PE configuration will

remain fixed in subsequent job segments as they are automatically submitted by the previous job segment. It is one of the goals of the Throughput Benchmark to simulate this aspect of GFDL's batch production environment.

Jobs may be run from existing executables. Time for compilation and linking as will be seen by the user of the delivered system will be reported elsewhere.

J.3.2.2.2 Running the LSC Throughput Benchmark

The Offeror will submit all of the 29 experiment first segments to the test queuing system in the following order:

1) FMS Spectral Atmosphere T42L20 coupled to 2deg MOM3 ocean
2) FMS Spectral Atmosphere T42L20 coupled to 1deg MOM3 ocean
3) FMS Spectral Atmosphere T106L30
4) FMS Lo-resolution N45L20 coupled to 1deg MOM3 ocean
5) FMS Hi-resolution N90L30 coupled to 1deg MOM3 ocean
6) FMS Standard Atmosphere N45L160
7) FMS Development N30L40
8) FMS Atmosphere N30L40 with Tracers
9) FMS Hi-resolution N270L40 Atmosphere
10) MOM3 2deg L36 ocean
11) MOM3 1deg L50 ocean
12) MOM3 3deg L25 + tracers
13) MOM3 p5deg MESO
14) HIM p25deg MESO
15) FMS Spectral Atmosphere T42L20 coupled to 2deg MOM3 ocean
16) FMS Spectral Atmosphere T42L20 coupled to 1deg MOM3 ocean
17) FMS Spectral Atmosphere T106L30
18) FMS Lo-resolution N45L20 coupled to 1deg MOM3 ocean
19) FMS Hi-resolution N90L30 coupled to 1deg MOM3 ocean
20) FMS Standard Atmosphere N45L160
21) FMS Atmosphere N30L40 with Tracers
22) MOM3 p5deg MESO
23) FMS Spectral Atmosphere T42L20 coupled to 2deg MOM3 ocean
24) FMS Spectral Atmosphere T42L20 coupled to 1deg MOM3 ocean
25) FMS Lo-resolution N45L20 coupled to 1deg MOM3 ocean
26) FMS Hi-resolution N90L30 coupled to 1deg MOM3 ocean
27) FMS Standard Atmosphere N45L160
28) FMS Spectral Atmosphere T42L20 coupled to 2deg MOM3 ocean
29) FMS Lo-resolution N45L20 coupled to 1deg MOM3 ocean

This submission may itself be performed through a shell script. The next segment of a job

stream will be submitted as part of the completion process of the segment which is running. The details of the submission scenario are described below.

The queuing system should be "live" and begin initiating jobs as they are submitted. The Government acknowledges that there may be start-up effects associated with flooding the queuing system with 29 jobs, but knows of no other reasonable way to assign a start time from which to measure the required 4 hour runtime maximum. The start time is measured from the time the first job is submitted (e.g., when the “Enter” key is pressed to execute the submission shell script).

It is desirable to run the Throughput Benchmark on a Test System that is as close to the offered system as possible. Offerors are cautioned that benchmark environments, procedures and methodologies which are judged by the Government to lack generality and/or extensibility for the lab will be penalized and run the risk of being rejected outright as non-compliant.

The details of the Throughput Benchmark job structure are as follows:

1. FMS Lo-resolution T42L20 coupled to 2deg MOM3 ocean

    a. 4 jobs of 2 segments each comprised of 180 days per segment (FMS: time_units=days; trun_length=180; MOM: NDAYS=180.0, diag=18.0).
    b. Segment 1 of each job is started from input data.
    c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
    d. **Segment 2 is re-run from the original input data; there is no unique segment 2**.

NOTE: Because there are multiple streams of this job running over the same time domain, care must be taken that output from one stream does not overwrite that of another. Moreover, the output from segment 1 should not overwrite the output from segment 2.

2. FMS Hi-resolution T42L20 coupled to 1deg MOM3 ocean

    a. 3 jobs of 2 segments each comprised of 21 days per segment (FMS: time_units=days; trun_length=21; MOM: NDAYS=21.0, diag=3.0).
    b. Segment 1 of each job is started from input data.
    c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
    d. **Segment 2 is re-run from the original input data; there is no unique segment 2**.

NOTE: Because there are multiple streams of this job running over the same time domain, care must be taken that output from one stream does not overwrite that of another. Moreover,

the output from segment 1 should not overwrite the output from segment 2.

3. FMS Spectral Atmosphere T106L30

   a. 2 jobs of 2 segments each comprised of 48 hours per segment (time_units=hours; trun_length=48).
   b. Segment 1 of each job is started from input data.
   c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
   d. Segment 2 is run from the input data and the restart file generated by the successful completion of segment 1.

NOTE: Because there are multiple streams of this job running over the same time domain, care must be taken that output from one stream does not overwrite that of another.

4. FMS Lo-resolution N45L20 coupled to 1deg MOM3 ocean

   a. 4 jobs of 2 segments each comprised of 14 days per segment: (FMS: time_units=days; trun_length=14; MOM: NDAYS=14.0, diag=2.0)
   b. Segment 1 of each job is started from input data.
   c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
   d. **Segment 2 is re-run from the original input data; there is no unique segment 2**.

NOTE: Because there are multiple streams of this job running over the same time domain, care must be taken that output from one stream does not overwrite that of another. Moreover, the output from segment 1 should not overwrite the output from segment 2.

5. FMS Hi-resolution N90L30 coupled to 1deg MOM3 ocean

   a. 3 jobs of 2 segments each comprised of 5 days per segment (FMS: time_units=days; trun_length=5; MOM: NDAYS=5.0, diag=1.0).
   b. Segment 1 of each job is started from input data.
   c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
   d. **Segment 2 is re-run from the original input data; there is no unique segment 2**.

NOTE: Because there are multiple streams of this job running over the same time domain, care must be taken that output from one stream does not overwrite that of another. Moreover, the output from segment 1 should not overwrite the output from segment 2.

6. FMS Standard Atmosphere N45L160

   a. 3 jobs of 2 segments each comprised of 45 hours per segment: (time_units=hours; trun_length=45)
   b. Segment 1 of each job is started from input data.
   c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
   d. Segment 2 is run from the input data and the restart file generated by the successful completion of segment 1.

NOTE: Because there are multiple streams of this job running over the same time domain, care must be taken that output from one stream does not overwrite that of another.

7. FMS Development N30L40

   a. 1 job of 2 segments comprised of 17 days per segment: (time_units=days; trun_length=17)
   b. Segment 1 of the job is started from input data.
   c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
   d. Segment 2 is run from the input data and the restart file generated by the successful completion of segment 1.

8. FMS Atmosphere N30L40 with Tracers

   a. 2 jobs of 2 segments each comprised of 17 days per segment: (time_units=days; trun_length=17)
   b. Segment 1 of each job is started from input data.
   c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
   d. Segment 2 is run from the input data and the restart file generated by the successful completion of segment 1.

NOTE: Because there are multiple streams of this job running over the same time domain, care must be taken that output from one stream does not overwrite that of another.

9. FMS Hi-resolution N270L40 Atmosphere

   a. 1 job of 2 segments comprised of 69 minutes per segment: (time_units=minutes; trun_length=69)
   b. Segment 1 of the job is started from input data.
   c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.

d. Segment 2 is run from the input data and the restart file generated by the successful completion of segment 1.

10. MOM3 1deg L50 ocean

a. 1 job of 2 segments comprised of 10.5 days per segment: (days=10.5, diag=1.75)
b. Segment 1 of the job is started from input data (initial=true).
c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
d. Segment 2 is run from the input data and the restart file generated by the successful completion of segment 1 (initial=false)

11. MOM3 2deg L36 ocean

a. 1 job of 2 segments comprised of 180 days per segment: (days=180.0, diag=30.0)
b. Segment 1 of the job is started from input data at t=0 (initial=true).
c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
d. Segment 2 is run from the input data and the restart file generated by the successful completion of segment 1 (initial=false).

12. MOM3 3deg L25 + tracers

a. 1 job of 2 segments comprised of 1800 days per segment: (days=1800.0, diag=300.0)
b. Segment 1 of the job is started from input data at t=0 (initial=true).
c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
d. Segment 2 is run from the input data and the restart file generated by the successful completion of segment 1 (initial=false).

13. MOM3 p5deg MESO

a. 2 jobs of 2 segments each comprised of 45 days per segment: (days=45.0, diag=7.5)
b. Segment 1 of each job is started from input data at t=0 (initial=true).
c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
d. Segment 2 is run from the input data and the restart file generated by the successful completion of segment 1 (initial=false).

NOTE: Because there are multiple streams of this job running over the same time domain, care must be taken that output from one stream does not overwrite that of another.

14. HIM p25deg MESO

a. 1 job of 2 segments comprised of 2592 time steps per segment.
b. Segment 1 of the job is started from input data at t=0 (initial=true).
c. After "storage" of all output files to archive disk, the segment 1 run script should submit the script for segment 2.
d. Segment 2 is simply a re-run of segment 1.

NOTE: The output from segment 1 should not overwrite the output from segment 2.

NOTE: Three of the job streams have been combined into a single experiment for HIM (that is the total run length of HIM has been specified such that it would take 36-T90 equivalent processor hours to complete). This will require that HIM be run on a proportionately larger number of processors than other experiments in the job mix.

As each experiment has 2 segments, there are a total of 29x2=58 experiment segments to be run for the Throughput Benchmark. There should be unique ASCII and archive output for each segment at the end of the throughput test.

As per section J.3.1.1, Source Code Changes, the baseline measurements required of all compliant offers must be made with only Class A modifications using MPI as the message passing library for those systems employing an explicit message communication library in the benchmark. Any extrapolations of values from Test Systems to the "baseline" performance of the offered system must be based on this data.

As further described in section J.3.1.1, the Offeror may supply additional measurements and extrapolations based on any combination of Class A, B, C, or D modifications. But as noted, such a data set is accepted and assessed risk, or rejected, as a whole. The Government will not attempt to selectively assess modifications associated with a given data set.

J.3.2.2.3 LSC Throughput Benchmark Output

The Offeror shall keep the responses to this section focused on the technical and engineering aspects of the benchmark data as pertains to their proposed solutions. Appropriate data includes CONCISE descriptions of Test System configuration and extrapolation and demonstration methodologies. References to competitors or other aspects of the general computing market place are NOT appropriate material for this section.

1. Provide a complete, concise description of the system configuration used for the

Throughput Benchmark. Be sure to include:

A. the queues and scheduling used to run the Throughput Benchmark
B. the job submission {environment, process and command lines}
C. all system operator activity during the runtime of the benchmark
D. the number of PEs on the Test System
E. the PE characteristics (e.g. processor cycle time and peak performance)
F. the cache configuration of each PE
G. the total and application memory available to each PE
H. the "communication fabric" of the system (where applicable)
I. the hardware and software supporting the file system(s) for the benchmark
J. how the archive disk is being simulated for the benchmark

2. Provide a complete, concise description of the data gathering procedures and the data gathered and the extrapolation methodology used. All timings are to be presented in whole units of seconds. Fractional timings which are less than 0.5 shall be rounded "down" to the nearest integer; timings which are greater than or equal to 0.5 shall be rounded "up" to the nearest integer.

3. With respect to the data provided in 1., how will the installed system differ from the Test System used for the RFP response? How does the data provided and the extrapolations from the Test System show that the installed system will perform as offered?

4. The file "LSC_Benchmarks.xls" has been distributed with the benchmark codes. In this file, an Excel 97 Throughput spreadsheet template has been provided for the Throughput Benchmark. One spreadsheet must be completed for each of the following cases:

   A. Running the Throughput Benchmark on the Test System with Class A modifications
   B. Running the Throughput Benchmark on the Test System with Class A-D modifications, if distinct from A.
   C. Running the Throughput Benchmark on the Offered system with Class A modifications, if distinct from A.
   D. Running the Throughput Benchmark on the Offered system with Class A-D modifications, if distinct from C.

5. Please return all verification files, cited in each benchmark's README file, that were produced on the Test System during the execution of the Throughput Benchmark.

The Government requires the following data be recorded in the Throughput spreadsheet for each experiment segment:

Column Heading	Definition
#PE	The number of PEs employed for the run
Run WCT	The wall clock time (WCT) from initiation to termination of the segment run script
Seg WCT	The WCT required from program invocation to program end for the segment
Agg CPU	The aggregate CPU time (user + system) used by the program
Agg Mem Use	The aggregate memory "highwater" mark
PE Mem Use	The per PE memory "highwater" mark

At the top of each spreadsheet, the end-to-end throughput wall clock time must be filled in.

J.3.2.3 LSC Scaling Study

J.3.2.3.1 General Comments

The goal of the Scaling Study is to measure individual application performance, scaling and resource requirements. Descriptions of the individual benchmark experiments are provided with each of the benchmark codes. See the README files included with the benchmark source for details. **Data for the Scaling Study should be collected using the same Test System that was used for the Throughput Benchmark**.

Applications should be run on as few processing elements as practical for the given experiment and should be scaled to as many PEs as possible. It is clear that at some number of PEs, the performance improvement of an application with respect to a particular experiment may flatten and perhaps decline. Termed a performance "rollover" point of the scaling curve, the Government requires data and documentation up to and including this point for all of the experiments.

The Government requires scaling data to 50% of the PEs on the offered system for the following experiments regardless of the presence of rollover points in the scaling curve:

9. FMS Hi-resolution N270L40 Atmosphere
14. HIM p25deg MESO

There may be multiple rollover points in the scaling curves for these experiments. Offerors may provide data beyond the first rollover point for other experiments at their discretion.

J.3.2.3.2 Running the LSC Scaling Study

In order to obtain a reasonable understanding of the scaling curve, the Government requires the following minimum number of performance data points for each experiment:

#	Experiment	Description	# data points
1	FMS Spectral Atmosphere T42L20 coupled to 2deg MOM3 ocean	1 segment (FMS: time_units=days; trun_length=30; MOM: NDAYS=30.0, diag=30.0)	5
2	FMS Spectral Atmosphere T42L20 coupled to 1deg MOM3 ocean	1 segment (FMS: time_units=hours; trun_length=84; MOM: NDAYS=3.5, diag=3.5)	5
3	FMS Spectral Atmosphere T106L30	1 segment (time_units=hours; trun_length=8)	5
4	FMS Lo-resolution N45L20 coupled to 1deg MOM3 ocean	1 segment (FMS: time_units=hours; trun_length=54; MOM: NDAYS=2.25, diag=2.25)	5
5	FMS Hi-resolution N90L30 coupled to 1deg MOM3 ocean	1 segment (FMS: time_units=hours; trun_length=18; MOM: NDAYS=0.75, diag=0.75)	5
6	FMS Standard Atmosphere N45L160	1 segment (time_units=minutes; trun_length=450)	5
7	FMS Development N30L40	1 segment (time_units=hours; trun_length=68)	4
8	FMS Atmosphere N30L40 with Tracers	1 segment (time_units=hours; trun_length=68)	4
9	FMS Hi-resolution N270L40 Atmosphere	1 segment (time_units=minutes; trun_length=12)	6
10	MOM3 1deg L50 ocean	1 segment (days=1.75, diag=1.75)	5
11	MOM3 2deg L36 ocean	1 segment (days=30.0, diag=30.0)	5
12	MOM3 3deg L25 + tracers	1 segment (days=300.0, diag=300.0)	4
13	MOM3 p5deg MESO	1 segment (days=7.5, diag=7.5)	5
14	HIM p25deg MESO	1st segment for 144 time steps	6

Run scripts for the scaling studies have been provided with the source code.

The Government requires that at least one of the data points be "reasonably close" (i.e. plus or minus 10%) to 1/30 of the proposed number of application PEs for the LSC for experiments 1-13 and 1/10 for experiment 14. It is acknowledged that experiment 12 is likely to scale poorly at 1/30 of the application PEs.

Data points should be provided at reasonable intervals between the minimum number of processors used and the maximum. As an example, a requirement for "6 data points" in an experiment which needs to span "minimum practical number of PEs" to "50% of the offered sytem" on a system with 1024 application PEs might look something like the set {16,32,64,128,256,512}. Offerors are encouraged to use processor configurations taking advantage of a "load balanced" number of PEs where this proves advantageous. Offerors are free to provide more data points at their discretion.

As per section J.3.1.1, Source Code Changes, the baseline measurements required of all compliant offers must be made with only Class A modifications using MPI as the message passing library for those systems employing an explicit message communication library in the benchmark. Any extrapolations of values from Test Systems to the "baseline" performance of the offered system must be based on this data.

As further described in section J.3.1.1, the Offeror may supply additional measurements and extrapolations based on any combination of Class A, B, C, or D modifications. But as noted, such a data set is accepted and assessed risk, or rejected, as a whole. The Government will not attempt to selectively assess modifications associated with a given data set.

J.3.2.3.3 LSC Scaling Study Output

The data to be gathered and returned with the Scaling Study benchmark is as follows:

1. Provide a complete, concise description of the system configuration used for the Scaling Study if different from the Test System used for the Throughput Benchmark. Be sure to include:

   A. the job submission {environment, process and command lines}
   B. the number of PEs on the Test System
   C. the PE characteristics (e.g. processor cycle time and peak performance)
   D. the cache configuration of each PE
   E. the total and application memory available to each PE
   F. the “communication fabric” of the system (where applicable)
   G. the hardware and software supporting the file system(s) for the benchmark
   H. how the archive disk is being simulated for the benchmark

2. Provide a complete, concise description of the data gathering procedures and the data gathered and the extrapolation methodology used. All timings are to be presented in

whole units of seconds. Fractional timings which are less than 0.5 shall be rounded "down" to the nearest integer; timings which are greater than or equal to 0.5 shall be rounded "up" to the nearest integer.

3. With respect to the data provided in 1., how will the installed system differ from the Test System used for the RFP response? How does the data provided and the extrapolations from the Test System show that the installed system will perform as offered?

4. The file "LSC_Benchmarks.xls" has been distributed with the benchmark codes. In this file, an Excel 97 Scaling Study spreadsheet template has been. One spreadsheet must be completed for each of the following cases:

   A. Running the Scaling Study on the Test System with Class A modifications
   B. Running the Scaling Study on the Test System with Class A-D modifications, if distinct from A.
   C. Running the Scaling Study on the Offered system with Class A modifications, if distinct from A.
   D. Running the Scaling Study on the Offered system with Class A-D modifications, if distinct from C.

   Items to be completed in the spreadsheet include:

      i. The time required to compile and link the application
      ii. The number of PEs employed for the run
      iii. The wall clock time from initiation to termination of the experiment run script
      iv. The wall clock time required from program invocation to program end for the experiment
      v. The aggregate CPU time used by the program
      vi. The per PE and aggregate memory "highwater" mark.

5. Please return all verification files, cited in each benchmark's README file, that were produced on the Test System during the execution of the Scaling Study.

J.3.3. ANALYSIS CLUSTER (AC) BENCHMARK

J.3.3.1 Overview

The AC benchmark is comprised of 2 parts with the following goals:

i) Throughput Benchmark: A measurement of system performance under quasi-realistic GFDL workload and Offeror proposed runtime environment.

il) Contention-Free Study: A measurement of individual application performance and resource requirements

There are 8 applications comprising the AC benchmark:

1. BASIN precipitation analysis
2. EIGEN eigenvalue calculation
3. SEASONAL postprocessing of climate integrations
4. LAN Analysis
5. LBL Line-by-Line radiation code
6. NC_COMBINE netcdf file combination
7. FMS Development N30L40
8. MOM3 3deg L25

These include unitasked applications as well as small parallel development codes (applications 7 and 8,whose source code has been distributed with the LSC benchmark) that users may wish to run on small (2-8) numbers of processors.

All experiments are to be run in 64-bit, IEEE floating point precision.

As with the LSC benchmark, the scenarios for RFP response and LTD systems apply. Similarly, the only acceptable post-award LTD will be to successfully run the entire throughput suite at the performance level proposed by the successful Offeror and as described by the Acceptance Criteria section of this document.

The Offeror may make changes to the application compilation and run scripts as necessary to accommodate their particular compilation and runtime environment(s).

Additionally, the Offeror may make changes to the source code. The same comments and classification scheme as described in Section J.3.1.1 applies to the AC benchmarks.

All requirements with regard to baseline performance measurements and evaluation of Class A modifications and Class A-D modifications are the same as for the LSC benchmarks as described in Section J.3.2.2 and will not be repeated here. Offerors are advised to review those instructions to ensure consistency between AC and LSC data.

Gathering of performance data is targeted to a system equivalent to that offered for the initial delivery. In this vein, it is highly desirable that the Test Systems used to provide performance numbers for the RFP response and the LTD be as close to the offered system as possible.

Still, the Government acknowledges that it may not be possible to use the offered system for either the RFP response or the LTD. See Section J.3.1.2. for details concerning RFP Test Systems and LTD systems.

J.3.3.2 AC Throughput Benchmark

All of the constraints for the LSC Throughput Benchmark apply to the AC Throughput Benchmark. A review of the constraints in Section J.3.2.2 would be prudent.

The AC Throughput Benchmark is comprised of a total of 40 job streams using the following number of job streams for each application (all job streams have one segment per stream):

1. BASIN precipitation analysis - 2 jobs
2. EIGEN eigenvalue calculation - 21 jobs
3. SEASONAL postprocessing of climate integrations - 2 jobs
4. LAN Analysis - 2 jobs
5. LBL Line-by-Line radiation code - 2 jobs
6. NC_COMBINE netcdf file combination - 4 jobs
7. FMS Development N30L40 - 3 jobs
8. MOM3 3deg L25 - 4 jobs

NOTE: When there are multiple streams of a job running at the same time, care must be taken that output from one stream does not overwrite that of another.

The small parallel applications (applications 7 and 8) should be run on 2 to 8 processors. Offerors should use resources adequate to produce performance consistent with that of the offered system.

The instructions for running the AC Throughput Benchmark are the same as for the LSC Throughput Benchmark described in Section J.3.2.2. All jobs in the AC throughput stream should be submitted to a "live" queuing system. Timing begins from submission of the first job. Although there is no specified time within which the AC Throughput Benchmark must complete, shorter completion times for the AC Throughput Benchmark will receive higher ratings in the evaluation.

The data to be gathered and returned for the AC Throughput Benchmark are identical to those to be gathered and returned for the LSC Throughput Benchmark, except the Throughput spreadsheet template to be used is found in the Excel 97 file “AC_Benchmarks.xls”, which has been distributed with the benchmark codes.

J.3.3.3 AC Contention-Free Performance Study

Descriptions of the individual benchmark experiments are provided with each of the benchmark codes. See the README files included with the benchmark source for details.

The goal of running the AC benchmark codes individually is to measure contention free application performance and resource requirements. Scaling of the parallel applications

7. FMS Development N30L40
8. MOM3 3deg L25

is not an issue. These jobs should be run on the same number of processors used for the jobs in the AC Throughput Benchmark.

The data to be gathered and returned for the AC Contention-Free Study are identical to those to be gathered and returned for the LSC Scaling Study, except the Contention-Free spreadsheet template to be used is found in the Excel 97 file “AC_Benchmarks.xls”, which has been distributed with the benchmark codes.

J.3.4. HIERARCHICAL STORAGE MANAGEMENT SYSTEM (HSMS) ARCHIVE BENCHMARK

J.3.4.1 Overview

The HSMS archive benchmark measures the sustained throughput for moving files between Analysis Cluster (AC) local scratch filesystem(s) and the HSMS.

The pre-award archive benchmark is designed to measure the performance of the network and protocols, or other interconnect, used to move files between the HSMS and the AC. Disk-to-disk file transfers are done between the AC and a computer which represents the HSMS. The complete HSMS software need not be used, but file transfer and filesystem software should be as close as possible to the offered system.

At installation, the archive benchmark must be run using the complete HSMS, including the nearline tier robotic library, under control of the HSMS software.

Both the pre-award and installation benchmarks must be run concurrently with the AC Throughput Benchmark and must complete in no more than 3600 seconds of wallclock time.

J.3.4.2 Running the Archive Benchmark

The archive benchmark is defined by (1) and (2) below. Approximately 48 dGB of data must be moved (24 dGB each way). In the pre-award benchmark, "the HSMS" means disk storage on a computer which represents the HSMS. In the installation benchmark, "the HSMS" means HSMS nearline tier tape storage as specified below. The AC local scratch filesystem(s) used must conform to the AC filesystem configuration proposed for production use.

(1) Move the specified number of copies of the files in the table below from the HSMS to AC local scratch filesystem(s):

(2) Move the specified number of copies of the files in the table below from AC local scratch filesystem(s) to the HSMS:

# copies	file size (dMB)	file
80	108	BASIN/archive/input/v2.precip.beta.1901_2000.unf3
20	783	LAN/archive/input/rthrm144

The files in the above table are analysis benchmark input files included in the analysis benchmark distribution. The requested copies should be prepared by running the provided "make_archive_files" script. This script creates files for the benchmark which follow an easily understood naming convention.

File transfers may be distributed over any combination of interactive and/or batch AC nodes. Batch nodes may be used without involvement of the batch queuing software. One possible implementation is a driver script run on one node which uses remote-shell commands to execute file transfers on several other nodes. The provided "run_in_parallel" script serves as an example driver script.

The benchmark execution time must be determined to the nearest second from "date" command output, AC process accounting reports, HSMS software log files, or other system timestamps.

In the installation benchmark, if tape technologies intended for small or large files are proposed, the 108 dMB files must be treated as small files, and the 783 dMB files must be treated as large files.

In the installation benchmark, for the files in (1) above which originate on HSMS tape storage, each file must reside on a separate tape volume. During setup of the installation benchmark, offerors must use administrator commands or other means to direct these files to separate tape volumes.

In the installation benchmark, execution time must include completion of writes to HSMS tape media. Also, the HSMS disk cache or staging filesystem must be cleared before running the benchmark, so that the files in (1) above are read entirely from tape storage.

J.3.4.3 Offeror Response

Offerors must provide written responses to the instructions below. In the technical proposal, respond for the pre-award benchmark only. At installation, respond for the installation benchmark. Reproduce each instruction above the response given.

1. Describe the hardware and software configuration used to run the benchmark, including file transfer and filesystem software.

2. Describe the distribution of the file transfers over the AC nodes. How does this differ from normal production use of the AC?

3. Give the benchmark execution time in seconds.

J.3.5 Legacy Archive Benchmark

J.3.5.1 Overview

The legacy archive benchmark measures the sustained throughput for moving files from the legacy archive to Analysis Cluster (AC) local scratch filesystem(s).

The legacy archive benchmark will be run stand-alone on the Analysis Cluster (AC) configured for production use. No other workload is run concurrently with this benchmark.

Throughout the base contract period, this benchmark must complete in no more than 1800 seconds of wallclock time. Higher levels of performance will not be given credit when evaluating proposals.

J.3.5.2 Running the Legacy Archive Benchmark

The legacy archive benchmark is defined by (3) below. Approximately 4.8 dGB of data must be moved from the legacy archive to AC local scratch filesystems(s).

(3) Move the files specified below from the legacy archive to AC local scratch filesystem(s):

# files	file size (dMB)	file names
32	50	legacy.S.01, ... legacy.S.32
4	783	legacy.L.01, ... legacy.L.04

These benchmark files will be prepared in the production legacy archive before HPCS installation. The 783 dMB files will be on 50 GB Redwood tapes, and the 50 dMB files will be on Timberline tapes. Each file will likely reside on a separate tape volume.

File transfers may be distributed over any combination of interactive and/or batch AC nodes. Batch nodes may be used without involvement of the batch queuing software. One possible implementation is a driver script run on one node which uses remote-shell commands to execute file transfers on several other nodes. The provided "run_in_parallel" script serves as

an example driver script.

The benchmark execution time must be determined to the nearest second from "date" command output, AC process accounting reports, or other system timestamps.

The legacy archive disk cache or staging filesystem must be cleared before running the benchmark, so that the files in (3) above are read entirely from tape storage.

for 512 MB and 1 GB of memory per processor are implemented, indicating if installed memory must be removed to perform the upgrade. Include the total memory on a node after each upgrade.

5. Node Cache Hierarchy. For nodes that run user applications, describe the cache hierarchy, including cache sizes in MB. What are the possible data paths between node memory and CPU registers? What are the hardware-level latencies and bandwidths for each part of each data path, with time given in CPU clock periods? What are the typical latencies and bandwidths when the effect of latency hiding is included?

6. Cluster Interconnection Network. Identify the products which make up the cluster interconnection network. Briefly describe the interconnection network. What are the features for data integrity checking and high availability? What is the bisection bandwidth of the proposed configuration?

7. Message Passing Within a Node. For multiple-CPU nodes that run user applications, give the latency and effective bandwidth of point to point and non-blocked bi-directional message passing within a node. Give values for raw message sizes of 2 bytes, 20 bytes, 200 bytes, 2 KB, 20 KB, 200 KB, and 2000 KB, or the closest available data points. The effective bandwidth b_e is defined as

$$b_e = \frac{N}{T_l + \left(N + N_E\right)/b}$$

where N is the raw message size, N_E is the size of the header and trailer information added to the raw message for purposes of transmission, T_l is the latency for sending the message, and b is the raw bandwidth. Describe the data paths taken and the sources of message-handling overhead. How do the available message passing interfaces differ in performance?

8. Message Passing Between Nodes. For nodes that run user applications, give the latency and effective bandwidth (as defined in instruction 7 of this tab) of point-to-point and non-blocked bi-directional message passing within a node. Give values for raw message sizes of 2 bytes, 20 bytes, 200 bytes, 2 KB, 20 KB, 200 KB, and 2000 KB, or the closest available data points. Describe the data paths taken and the sources of message-handling overhead. How do the available message-passing interfaces differ in performance?

9. LSC disk. Identify the numbers, types, and placement of disk subsystems, channels, and channel switches which provide the LSC disk storage. What type of fault-tolerant disk subsystem is proposed? How is performance impacted by failure of a disk drive? How many concurrent, independent disk I/O operations can be performed, to handle multiple jobs and interactive users? Show a calculation of the total formatted capacity in dTB, exclusive of RAID parity disks and other system use. Show a calculation of the total sustained I/O bandwidth of the proposed configuration. If zoned disks with a varying number of sectors-per-track are used, give two calculations,

2. Computers. Section C defines a computer to be “the maximum set of nodes that may be unavailable during the repair of any subset of those nodes.” Following this definition, identify the computers in the AC.

3. Nodes. Give the total number of nodes, and its breakdown into different node types. For nodes that run user applications, are all processors binary compatible, and do all nodes have identical hardware configurations? If not, describe the differences between the processors or nodes. What is the maximum number of nodes a single MPI message-passing application may use on the AC?

4. Node Configuration. For each type of node, give the number of CPUs in a node, the CPU type, clock speed, and byte order (big- or little-endian), the total memory in GB on a node, the node memory per CPU, the largest logically-shared address space on a node, the number and type of I/O channels on a node, and the amount of local disk in dGB on a node.

5. Node Cache Hierarchy. For nodes that run user applications, describe the cache hierarchy, including cache sizes in MB. What are the possible data paths between node memory and CPU registers? What are the hardware-level latencies and bandwidths for each part of each data path, with time given in CPU clock periods? What are the typical latencies and bandwidths when the effect of latency hiding is included?

6. Cluster Interconnection Network. Identify the products which make up the cluster interconnection network. Briefly describe the interconnection network. What are the features for data integrity checking and high availability? What is the bisection bandwidth of the proposed configuration?

7. Message Passing Within a Node. For multiple-CPU nodes that run user applications, give the latency and effective bandwidth of point to point and non-blocked bi-directional message passing within a node. Give values for raw message sizes of 2 bytes, 20 bytes, 200 bytes, 2 KB, 20 KB, 200 KB, and 2000 KB, or the closest available data points. The effective bandwidth b_e is defined as

$$b_e = \frac{N}{T_l + \left(N + N_E\right)/b}$$

where N is the raw message size, N_E is the size of the header and trailer information added to the raw message for purposes of transmission, T_l is the latency for sending the message, and b is the raw bandwidth. Describe the data paths taken and the sources of message-handling overhead. How do the available message passing interfaces differ in performance?

8. Message Passing Between Nodes. For nodes that run user applications, give the latency and effective bandwidth (as defined in instruction 7 of this tab) of point-to-point and non-blocked bi-directional message passing within a node. Give values for raw message sizes of 2 bytes, 20 bytes, 200 bytes, 2 KB, 20 KB, 200 KB, and 2000 KB, or

the closest available data points.